DOCUMENT RESUME

ED 452 209 TM 032 494

AUTHOR	Tanguma, Jesus
TITLE	Effects of Sample Size on the Power of Selected Fit Indices: A Graphical Approach.
PUB DATE	2001-02-00
NOTE	22p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, February 1-3, 2001). Some color figures may not reproduce well.
PUB TYPE	Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Goodness of Fit; Models; Monte Carlo Methods; *Sample Size; Simulation
IDENTIFIERS	*Power (Statistics)

ABSTRACT

The purpose of this study was to investigate the effects of sample size on the power of five selected fit indices through a Monte Carlo simulation. Two models (a reduced and a complete model) and 6 sample sizes (20, 50, 100, 200, 500, and 1,000) were used to investigate the effect on the power of fit indices as the sample size was varied. The power of the selected fit indices, more often than not, was different across sample sizes, indicating that sample size does affect the power of the fit indices. The results of this study indicate that of all the indices examined, the Goodness of Fit Index (GFI) was the most powerful fit index. (Contains 2 tables, 4 figures, and 26 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

TM

Running head: POWER OF FIT INDICES

ED 452 209

Effects of Sample Size on the Power of Selected Fit Indices: A Graphical Approach

Jesus Tanguma

University of Houston Clear Lake

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Jesus Tanguma

TM032494

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 2001.

BEST COPY AVAILABLE

Abstract

The purpose of the present study was to investigate the effects of sample size on the power of selected fit indices. Two models (i.e., a reduced and a complete model) and six (20, 50, 100, 200, 500, 1000) sample sizes were used to investigate the effect on the power of the fit indices as sample size was varied. The power of the selected fit indices, more often than not, was different across sample sizes, thus indicating that sample size does affect the power of the fit indices. The results of the present study indicated that of all the indices examined, GFI was the most powerful fit index.

Effects of Sample Size on the Power of Selected Fit Indices: A Graphical Approach

Structural Equation Modeling (SEM) is a comprehensive statistical approach used by researchers in education, psychology, sociology, econometrics, and other social sciences (Thompson, 2000). SEM (a) directly incorporates explicit estimation of measurement error (i.e., score reliability) and (b) is especially useful for addressing questions of score validity because theoretical models are directly tested. According to Gerbing and Anderson (1993), "the empirical assessment of proposed models is a vital aspect of the theory development process, and central to this assessment are the values of goodness-of-fit indices obtained from the analysis of a specified model" (p. 40).

Although more than 30 goodness-of-fit indexes have been reported and their empirical behavior has been studied (e.g., Marsh, Balla, & McDonald, 1988), there is no consensus among researchers as to which is the "best fit index" (Thompson & Daniel, 1996). Thus, "investigators may have difficulty choosing among" (Tanaka, 1993, p. 10) the existing fit indices. Some of the problems faced by researchers when evaluating model fit is that existing indices estimate no known population parameters (Bentler, 1990) and "measure misspecification at the level of covariances, and not at the level of the relevant structural parameters" (Saris & Satorra, 1993, p. 181). Another problem is that all goodness-of-fit indices, to some degree, are dependent on sample size.

For example, in their analysis of the more than 30 indexes Marsh et al. (1988) concluded that the "Tucker-Lewis index was the only widely used index that was relatively independent of sample size" (p. 391). Similar results have been reported by

other researchers (e.g., Bentler, 1990; Bollen, 1990; Fan, Thompson, & Wang, 1999; Fan, Wang, & Thompson, 1997; Hoelter, 1983; Mulaik, James, Alstine, Bennett, Lind, & Stilwell, 1989).

Although there is no consensus as to which is the "best fit index" (Thompson, 2000), Gerbing and Anderson (1993) suggested that the ideal goodness-of-fit index should

> (1) indicate degree of fit along a continuum bounded by values such as 0 and 1, where 0 reflects a complete lack of fit and 1 reflects perfect fit;
>
> (2) be independent of sample size (higher or lower values would not be obtained simply because the sample size is large or small); and
>
> (3) have known distributional characteristics to assist interpretation and allow the construction of a confidence interval. (p. 41)

However, no existing fit index satisfies all these ideal conditions.

A common practice among researchers when performing SEM analysis is to compare and evaluate several alternative models. This is because, as Thompson (2000) explained in the very first of his 10 commandments of good structural equation modeling behavior,

> 1. Never conclude that a model has been definitely proven, because infinitely many models can fit any given data set (thus, the fit of a single tested model is always an artifact of having tested too few models). (pp. 277-278)

For example, two competing models may differ by the direction of a path, the omission of a path, or the omission of one or more variables. In evaluating the various competing models, researchers may use the amount of variance explained in dependent variables,

size of regression coefficients, residuals, or goodness-of-fit indices, among others (Biddle & Marlin, 1987).

Because there is no consensus among the researchers as to which is the "best" fit index, "the analysis of the power of the chi-square test can be a very useful aid in assessing model fit" (Bollen, 1989, p. 349). Without knowledge of the power of the test (i.e., the probability of rejecting the null hypothesis when it is false), researchers cannot predict whether wrongly specified models may be rejected in small sample studies, while in large sample studies even minimal errors may lead to rejection of the model (Satorra & Saris, 1985).

The purpose of this study was to graphically investigate the power of some of the most commonly used fit indices (e.g., NFI, CFI, GFI, AGFI, and chi-square) varying sample size. That is, is the power of the selected fit indices similar or different for each sample size studied?

Although other researchers have discussed the power of fit indices, the approach taken in the present study was quite different from the rest. That is, whereas other researchers have investigated power from tables and histograms (e.g., Saris & Satorra, 1993; Satorra & Saris, 1985; Satorra, 1989; Saris, den Ronden, & Satorra, 1987; Marsh, Balla & McDonald, 1989), the approach taken here was to investigate the cumulative distribution of fit indices graphically as well as numerically. The recently-released report of the APA Task Force on Statistical Inference has placed an emphatic emphasis on the importance of using graphical techniques to explore and understand data (Wilkinson & APA Task Force on Statistical Inference, 1999). As the Task Force emphasized,

> As soon as you have collected your data, before you compute any statistics, look at your data. Data screening is not data snooping. It is not an opportunity to discard or change values to favor your hypotheses. However, if you assess hypotheses without examining your data, you risk publishing nonsense.... Graphical inspection of data offers an excellent possibility for detecting serious comprises to data integrity. The reason is simple: Graphics broadcast; statistics narrowcast. (p. 597, emphasis in original)

Certainly such admonitions are not new (Tukey, 1977; Wilkinson, 1999).

Method

A Monte Carlo simulation approach was taken to investigate the power of the goodness-of-fit indices. The model for investigation was based on research by Brossart, Willson, Patton, Kivlighan, and Multon (1999) of a counselor-client interaction. A pictorial representation of the model is presented in Figure 1. Parameters with the same subscripts are restricted to have the same value. Each realization consisted of 20 time points.

After deleting the paths from CO1 to CL3 and from CO2 to CL3, the reduced (less restrictive) model is obtained. The dashed line/paths in Figure 1 indicate a removed path under the reduced model. By deleting these paths, their effects are assumed to be zero.

Simulation for the baseline and reduced models was developed using SAS for PC (SAS Institute, 1989). From extension of the work of Kim (1999), a macro was developed to randomly generate 200 simulations for each condition for both "baseline" and "reduced" models being investigated. Sample size was varied from 20, 50, 100, 200,

500, and 1000 replications. Each replication in each simulation was analyzed using PROC CALIS, under both the "baseline" and "reduced"model conditions. Each replication consisted of 20 time points for CO and CL fit to the lag autoregressive process represented in the baseline model and the reduced model.

Once the data had been generated, they were imported into SPSS (SPSS Inc., 1999). All subsequent analyses were done using SPSS 10.0. To detect the effects of sample size on the power of the selected fit indices, an ogive of the distribution of each fit index per sample size was constructed using an SPSS procedure. Tanguma and Speed (2000) reported the logic used to develop these ogive graphs.

Results

The current study employed two models (baseline and reduced) and six (20, 50, 100, 200, 500, and 1000) sample sizes to investigate the effect on the power of the fit indices (GFI, AGFI, CFI, NFI, and chi-square) as sample size varied. Several previous studies have examined the effect of sample size on fit indices and have presented their findings in the form of tables and histograms. However, although the findings of the present study concerning the effects of sample size on fit indices are consistent with the literature, the findings are presented using tables, histograms, and ogive plots. The use of ogive plots enhances the researcher's visual perception of the impact of sample size on the fit indices (Wilkinson & APA Task Force on Statistical Inference, 1999).

Power of the Fit Indices

When doing structural equations modeling, researchers should keep in mind that their decision to reject or fail to reject a given model should not be based solely on fit indices. After all, all fit indices depend on sample size to some degree. The power of

statistics test should also be considered. The power of a test is defined as the probability of rejecting an incorrect model. According to Saris and Satorra (1993), a procedure to calculate the power of the likelihood ratio test in structural equations modeling is

$$\pi = \Pr[\chi^2_{df}(\lambda) > c_\alpha]$$

where the noncentrality parameter λ may be computed according to several procedures. The null and alternative hypotheses for a power analysis represent two models, one nested within the other by constraining one or more parameters in the first model. For example, according to Saris and Satorra (1993), the noncentrality parameter may be computed as follows:

$$\lambda = Min_{\theta \in H_0} nF\left[\sum(\theta_A), \sum(\theta)\right].$$

Notice that in computing the noncentrality parameter, the model is fitted using the original parameter value (θ) and the alternative parameter value (θ_A).

The power of the test may be evaluated one parameter at a time or several parameters at once. That is, the effect on power of including or omitting one or several parameters may be tested at once.

A review of the literature has shown that although other researchers (e.g., Satorra & Saris, 1985; Saris & Stronkhorst, 1984; Satorra, 1989; Satorra, Saris, & de Pijper, 1991, and Saris & Satorra, 1993) have investigated the power of the test of the likelihood ratio statistic, no published research has been done on the power of other fit indices. Thus, a procedure to estimate power graphically for goodness-of-fit indices was developed in the present study. This procedure compares two models: a complete and a reduced model. These comparisons are done via tables, histograms, and ogive plots.

The power of a given fit index at a specific sample size is computed in several steps. First, for a given sample size of the reduced model, the value of the fit index at the 95th percentile is identified. This value is then used as a cut off point in the distribution of values for the complete model. The total number of values at or beyond this cut off point is determined. This number is then divided by the number of values in the distribution. The result is defined as the power of the fit index. For example, to determine the power of GFI when sample size is 20, the respective 95th percentile for the reduced model is found to be 0.955. Then, the number of values in the distribution, for the complete model, which are greater than or equal to 0.955 are counted. In this example there are 9 such values. Next, the proportion of values at or beyond the cut off point to the total number of values is computed, 9/200 = 0.045. This value is defined as the power of the GFI when n = 20. Similarly, when computing the power of CFI when n = 500, it is determined that there are 30 values in the distribution of the complete model which are at or beyond the 95th percentile in the reduced model. Thus, the power of CFI at n = 500 is 30/200 = 0.150. Table 1 lists the results of computing the power of each fit index at the different sample sizes.

Insert Table 1 About Here

Another way to determine the power of a given fit index is to graph the ogives of the reduced and complete models for a given sample size. For example, looking at the ogive plot for CFI when n = 500 (see Figure 2), it is obvious that a large number (30) of values are at or beyond the reduced model's 95th percentile. Similarly, looking at the ogive for chi-square when n= 1000 (see Figure 3), one can see that very few (5) of the

values in the complete model's distribution are at or beyond the reduced model's 95th percentile. Thus, the power of the CFI (0.150) when n = 500 is much larger than the power of chi-square (0.025) when n = 1000.

Insert Figures 2 and 3 About Here

Discussion

The dependency of the fit indices on sample size has forced researchers to search for other methods of evaluating the fit of the model. One such method is to look at the power of the test. That is, are researchers rejecting what they wanted to reject? Said differently, are researchers rejecting the null hypothesis when it is in fact false?

The results of the power analysis for each fit index are presented in Table 2. A graphical representation of the results of computing the power of each fit index at the different sample sizes is shown in Figure 4. Notice how for each sample size studied, the power of the selected fit indices varied. For example, the power analysis for CFI indicated that only two (n = 50 and n = 200) of the six sample sizes had equal power (0.090). Similarly, when n= 20 and again when n =1000 the power for NFI was the same (0.075).

Insert Figure 4 About Here

Insert Table 2 About Here

As depicted in Table 2, four out of six times AGFI had the lowest power values of all the fit indices. Thus, it had the lowest mean power value across all fit indices and across all sample sizes.

For sample sizes less than 100, NFI was the fit index that had the highest power on both occasions. Similarly, GFI was the fit index with the highest power when n = 500 and again when n = 1000. Only when n = 200 was CFI the fit index with the highest power.

Of the six fit indices investigated, the goodness-of-fit index (GFI), on the average, had the highest power, followed by the comparative fit index (CFI) and the normed fit index (NFI), with the others trailing behind. Similarly, the adjusted-goodness-of-fit index (AGFI) and chi-square (χ^2)—two commonly used indices—performed less well.

Limitations

In this study, only five of the more than 30 goodness-of-fit indices were considered. Thus, no statements can be made as to how the cumulative distribution of other fit indices might be affected by varying sample size. Similarly, only six sample sizes were used in the study. Consequently, no statements can be made as to how the cumulative distributions of the fit indices may be affected by sample sizes other than those in the study. Also, the deletion of a different path than the one deleted in this study may have different effects on the distribution of the fit indices. The fit indices analyzed in this study are commonly outputted by software packages such as AMOS and SAS PROC CALIS, among others. The cumulative distribution for each of the fit indices was analyzed using tables, histograms, and ogive plots. However, it would be useful to extend to additional fit indices, and especially the root mean square residual and the root mean square error of approximation.

Recommendations

In the future, it may be useful to examine the effect on the cumulative distribution of the fit indices when a different path is deleted. It may also be instructive to study the cumulative distribution of the fit indices when other sample sizes are used. Similarly, it may be useful to study the effects on power when other sample sizes are used.

Generally, sample size does affect (to some extent) the power of all fit indices. However, for a given model, the degree to which the power of the different fit indices are affected varied from sample size to sample size. For example, it was determined that the power of AGFI was the most affected and that of GFI was the least affected as sample size was varied from 20 to 1000.

The power of the selected fit indices, more often than not, was different across sample sizes, thus indicating that sample size also affects the power of the fit indices. The results of this study indicated that of all the indices examined, GFI was the most powerful fit index.

References

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107, 238-246.

Biddle, B. J., & Marlin, M. M. (1987). Casuality, confirmation, credulity, and structural equation modeling. Child Development, 58, 4-17.

Bollen, K. A. (1989). Structural equations with latent variables. New York: Wiley.

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects. Psychological Bulletin, 107, 256-259.

Brossart, D., Willson, V., Patton, M., Kivlighan, Jr., D., & Multon, K. (1998). A time series model of the working alliance: A key process in short-term psychoanalytic counseling. Psychotherapy, 35, 197-205.

Fan, X., Thompson, B., & Wang, L. (1999). The effects of sample size, estimation methods, and model specification on SEM fit indices. Structural Equation Modeling, 6, 56-83.

Fan, X., Wang, L., & Thompson, B. (1997). Effects of data nonnormality on fit indices and parameter estimates for true and misspecified SEM models. Paper presented at the annual meeting of the American Educational Research Association, Chicago. (ERIC Document Reproduction Service No. ED 408 299)

Gerbing, D. W., & Anderson, J. C. (1993). Monte Carlo evaluations of goodness-of-fit indices for structural equation models. In K. A. Bollen & J. S. Long (Eds.), Testing structural equation models (pp. 40-65). Newbury Park, CA: Sage.

Hoelter, J. W. (1983). The analysis of covariance structures: Goodness-of-fit indices. Sociological Methods & Research, 2, 325-344.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

Mulaik, S. A., James, L. R., Alstine, J. V., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105, 430-445.

Saris, W. E., den Ronden, J., & Satorra, A. (1987). Testing structural equation models. In P. F. Cuttance, & J. R. Ecob (Eds.), Structural modeling. Cambridge: Cambridge University Press.

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. Bollen., & J. S. Long (Eds.), Testing structural equation models (pp.181-204). Newbury Park, CA: Sage.

Saris, W. E., & Stronkhorst, L. H. (1984). Causal modeling in nonexperimental research: An introduction to the LISREL approach. Amsterdam: Sociometric Research Foundation.

SAS Institute, (1989). SAS user's guide: Statistics, version 6.01. Cary, NC: SAS Institute.

Satorra, A. (1989). Alternative test criteria in covariance structure analysis: A unified approach. Psychometrika, 54, 131-151.

Satorra, A., & Saris, W. E. (1985). Power of the likelihood ratio test in covariance structure analysis. Psychometrika, 50, 83-90.

Satorra, A., Saris, W.E., & de Pijper, W.M. (1991). A comparison of several approximations to the power function of the likelihood ratio test in covariance structure analysis. Statistica Neerlandica 45, 173-185.

SPSS. (1999). SPSS base 10.0. Chicago: Author.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. Bollen., & J. S. Long (Eds.), Testing structural equation models (pp.10-39). Newbury Park, CA: Sage.

Tanguma, J., & Speed, F.M. (2000). Graphing cumulative distribution functions. Educational and Psychological Measurement, 60, 488-497.

Thompson, B. (2000). Ten commandments of structural equation modeling. In L. Grimm & P. Yarnold (Eds.), Reading and understanding more multivariate statistics (pp. 261-284). Washington, DC: American Psychological Association.

Thompson, B. & Daniel, L.G. (1996). Factor analytic evidence for the construct validity of scores: An historical overview and some guidelines. Educational and Psychological Measurement, 56, 213-224.

Tukey, J. (1997). Exploratory data analysis. Cambridge, MA: Addison-Wesley.

Wilkinson, L. (1999). Graphs for research in counseling psychology. The Counseling Psychologist, 27, 384-407.

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604. [reprint available through the APA Home Page: http://www.apa.org/journals/amp/amp548594.html]

Table 1
Power analysis for fit indices

	GFI		AGFI		CFI		NFI		Chi-square	
n	95th %tile	Power	95th %tile	Power	95th %tile	Power	95th %tile	Power	95th %tile	Power
20	.955	.045	.905	.025	.975	.050	.971	.075	207.600	.030
50	.948	.085	.890	.035	.970	.090	.968	.100	417.700	.025
100	.946	.055	.887	.010	.968	.045	.968	.045	777.100	.055
200	.942	.065	.942	.050	.966	.090	.966	.070	1460.000	.035
500	.938	.175	.870	.005	.964	.150	.964	.135	3450.000	.045
1000	.937	.170	.869	.000	.964	.080	.964	.075	6800.178	.025

Table 2
Fit indices power analysis
Various sample sizes

Index	20	50	100	200	500	1000
GFI	0.045	0.085	0.055	0.065	0.175	0.170
AGFI	0.025	0.035	0.010	0.050	0.005	0.000
CFI	0.050	0.090	0.045	0.090	0.150	0.080
NFI	0.075	0.100	0.045	0.070	0.135	0.075
χ^2	0.030	0.025	0.055	0.035	0.045	0.025

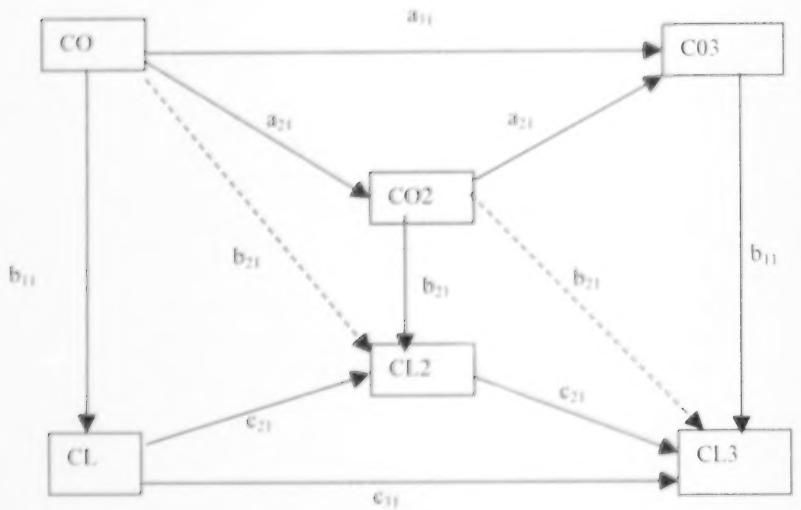

Figure 1 Counselor-client interaction model.
Note. CO1 = counselor working alliance score at any time; CO2 = counselor working alliance score at any time + 1; CO3 = counselor working alliance score at any time + 2; CL1 = client working alliance score at any time; CL2 = client working alliance score at any time + 1; CL3 = client working alliance score at any time + 2.

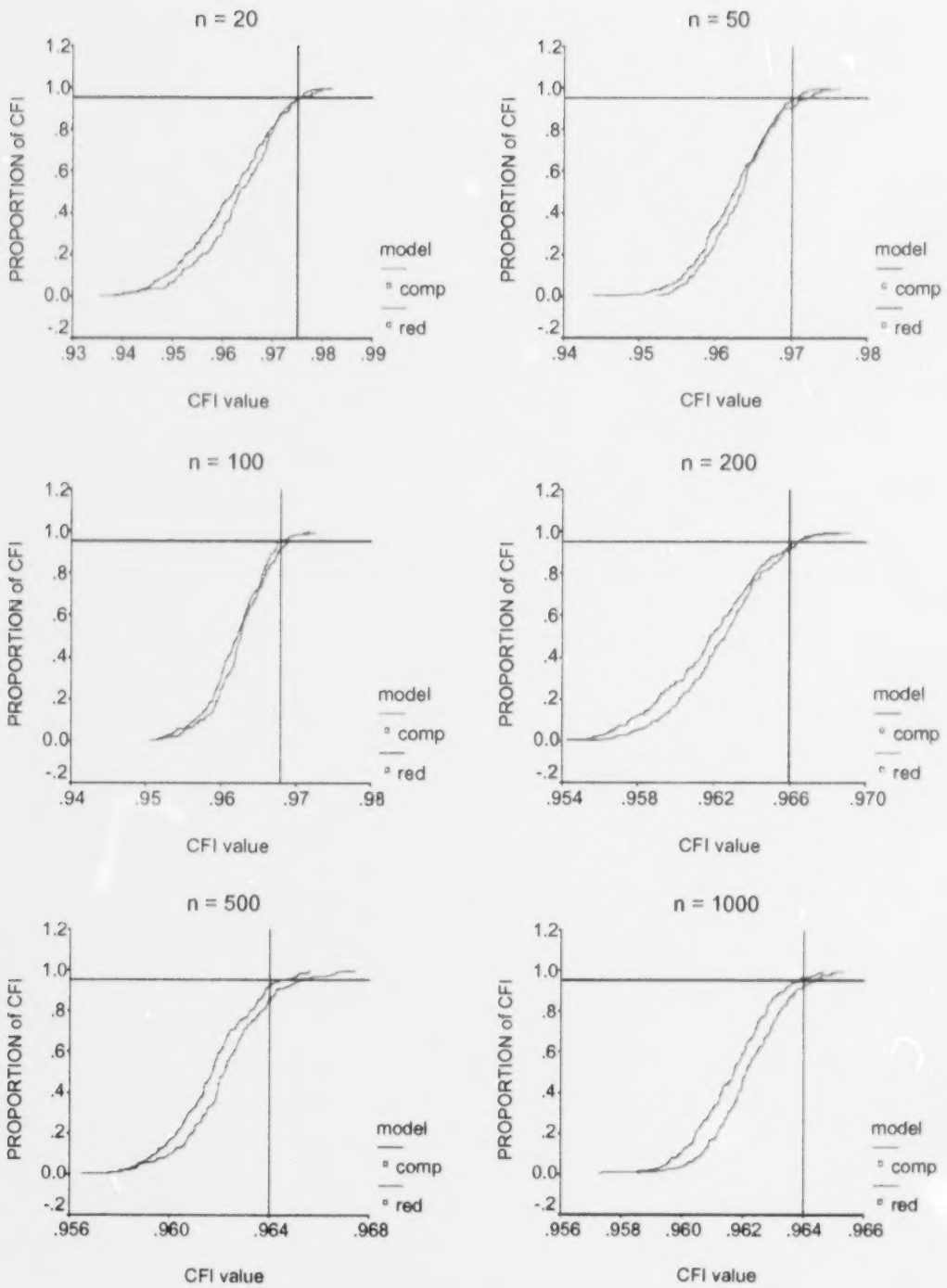

Figure 2 Power analysis for CFI at various sample sizes.

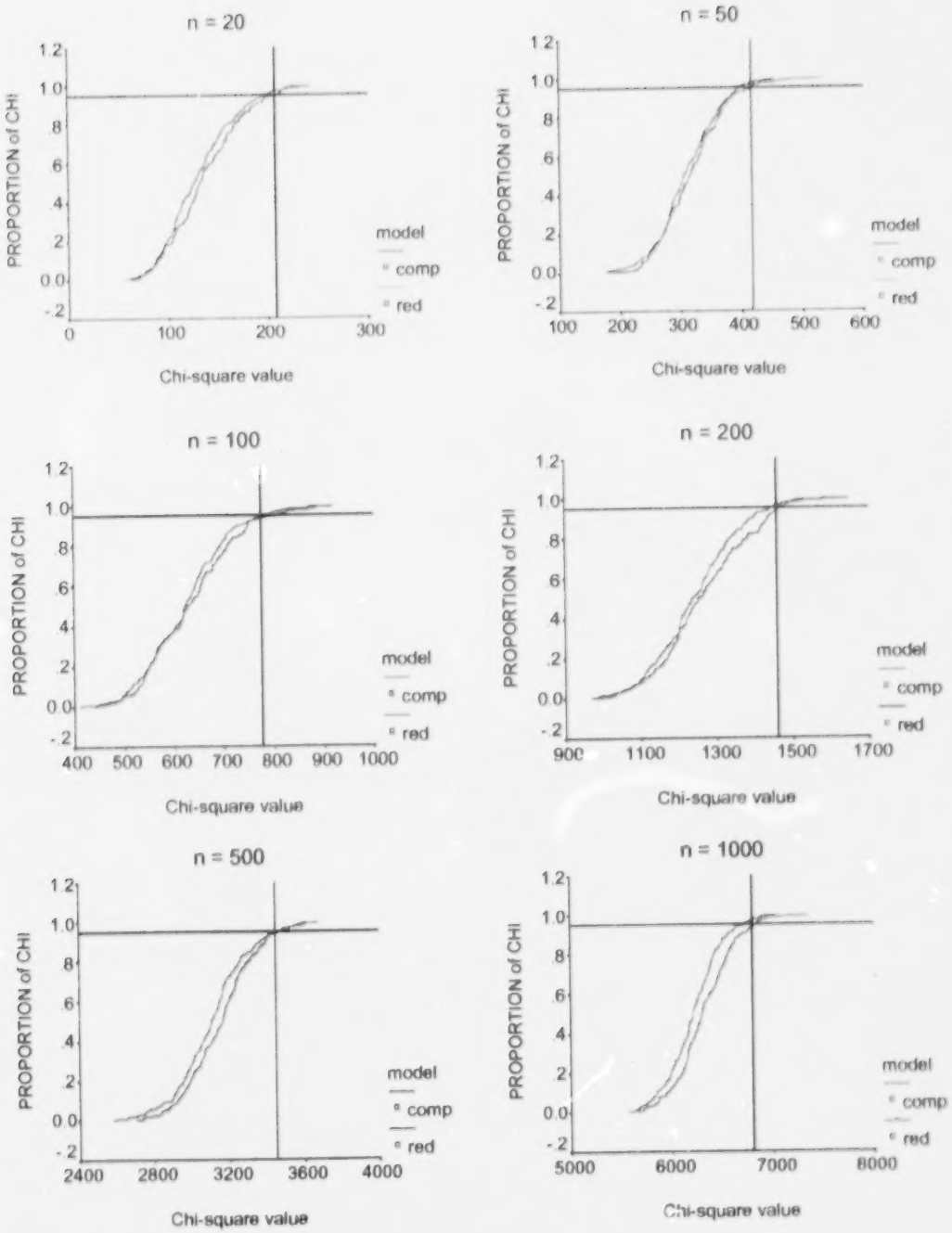

Figure 3 Power analysis for Chi-square at various sample sizes.

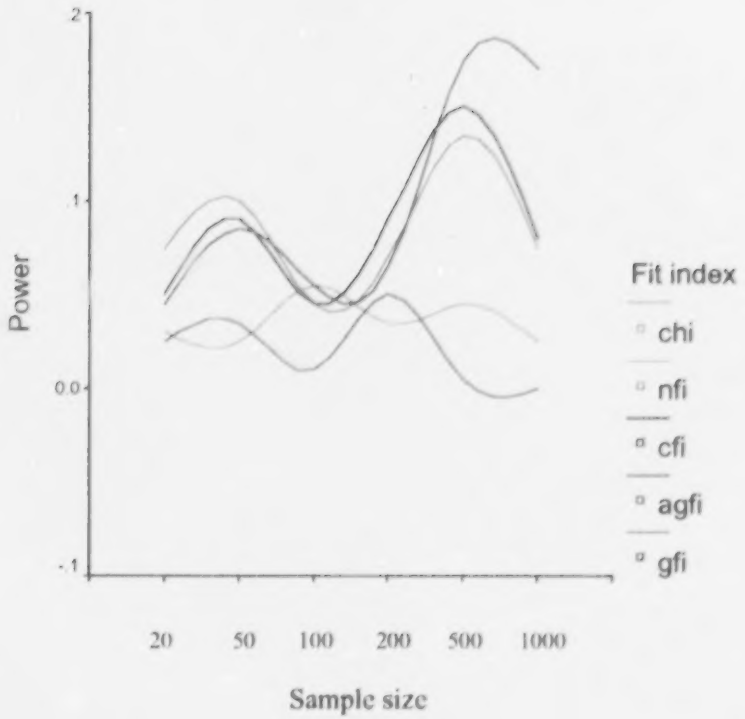

Figure 4 Power analysis of fit indices.